Retrieving Research Studies: A Comparison of Bibliographic and Full-Text Versions of the New England Journal of Medicine

E. Diane Johnson**, M.A., MaryEllen C. Sievert*, Ph.D., Emma Jean McKinin*, M.A.

**J. Otto Lottes Health Sciences Library
*School of Library and Informational Science
University of Missouri-Columbia, Columbia, MO 65211

It has been established that subject searches of medical full-text databases obtain higher recall than subject searches in a bibliographic database. In this study we attempted to determine if the same rule might apply when searching for a non-subject parameter such as study design. A simultaneous search of bibliographic and full-text records from the New England Journal of Medicine provided data on the number of items retrieved by each kind of search. Filtering strategies were created for 5 different study types: randomized controlled trials, other clinical trials and prospective studies, cohort studies, longitudinal and follow-up studies, and multicenter studies. The point of the study was to compare the numbers of items retrieved from the bibliographic database, MEDLINE, and those retrieved from the full-text version of NEJM, and to examine the unique access points available in each file. For all the study types the full-text file retrieved a larger number of records than MEDLINE, most of which were retrieved because of methodology terms found in the text but not in the title or abstract. In MEDLINE, descriptors and publication types, two value-added fields supplied by indexers, retrieved 11-89% more than title and abstract alone.

INTRODUCTION

In the last decade, there was much concern about access to methodologically rigorous studies from MEDLINE. Poor recall has been attributed to failures in indexing and the lack of appropriate search terms [1-4]. Because of a lack of confidence in the ability of MEDLINE to identify randomized controlled trials, it is not uncommon for those who maintain registries of clinical trials to resort to hand searches of journals in an attempt to identify all possible randomized trials [5]. It has been estimated that it takes 10 times longer to identify an eligible study through a manual search as compared to a computerized (MEDLINE) search [6]. Consequently, any time that a computerized search of the full-text can be substituted for a manual search there will be a substantial saving in time and money. Recent enhancements to MEDLINE, including the introduction of publication types and new MeSH terms to describe research designs, have improved the ability to retrieve methodologically rigorous studies [7]. In full-text databases every word in the article is searchable which eliminates the possibility of indexing failures, but raises the problem of synonymy. Until now, however, no one has examined how many additional citations could be retrieved by searching for methodology terms in the full-text. The MEDLINE/Full-Text Project is addressing this question. For the study discussed here we constructed search strategies for a variety of study types for both the bibliographic and the full-text files. In this way we were able to compile data on the number retrieved from each type of file, and from which fields they were retrieved.

Presented here are data for the following study types, which have all been recognized as attributes of research studies of varying degrees of rigor: randomized controlled trials, other clinical trials and prospective studies, multicenter studies, cohort studies, and longitudinal and followup studies. No attempt was made to apply the filters to a topic or to make relevance judgments on the sets retrieved. While we have studied the differences between bibliographic and the full-text databases of journal articles as part of the MEDLINE/Full-Text Research Project, we have never examined any advantage searching the full-text might have for retrieving a non-subject parameter such as study design. The point of the study was to compare the numbers of items retrieved from the bibliographic database, MEDLINE, and those retrieved from a full-text version of the New England Journal of Medicine (NEJM). We were then able to determine how many records in each database were retrieved only by fields unique to that file.

0195-4210/95/$5.00

METHODS

For this project we searched two databases available on the DIALOG search service (from Knight-Ridder Information, Inc.): MEDLINE and a database containing the full-text of NEJM from 1985 to date. Both databases were searched simultaneously using the Dialog OneSearch feature so that duplicates within each file could be detected and removed. At the outset, we believed that duplicate detection would also allow us to examine overlap between the two databases, but further testing revealed that DIALOG's duplicate detection algorithm did not work properly on MEDLINE items which contained a title rubric, such as "[see comments]".

The search of the two files MEDLINE (file 154) and NEJM (file 444) was limited to the 1990-1994 articles in the New England Journal of Medicine. We began with 1990 because that would allow us to use the publication types on MEDLINE and we ended with 1994 so we could eliminate the problem of current issues on one database not yet available on the other. Book reviews were eliminated from the full-text NEJM output.

The next step was to enter the term(s) for each filter. (A description of the strategies appears in the Appendix.) The terms were searched in the basic index. This insured that for the MEDLINE NEJM search the terms would be retrieved from: the title, abstract, descriptor, check tag and identifier fields. In the full-text database a basic index search covers the following fields: title, abstract, text (body of the article) and cited references.

Preliminary testing showed that there were duplicate records in both databases. That is, a single article might have been entered in one of the files two or three times instead of a single time. We used the DIALOG software command "rd" to remove duplicates from each database.

Next we did searches to determine how many records were retrieved owing to the searchable fields unique to each database. Since the title and abstract fields are searchable in both the bibliographic and full-text versions, they were of less interest than fields like descriptor and publication type (available only in MEDLINE) or text and cited references (available only in the full-text file).

For each filter, we searched to determine how many of the records were retrieved because of the presence of search terms in the "value-added" MEDLINE fields supplied by the indexer. Depending upon the filter being searched, this could include a publication type and/or a descriptor. From this set, we used the NOT operator to remove items where the terms also appeared in the title or the abstract. The remaining MEDLINE records were only retrieved because of the presence of a descriptor and/or a publication type assigned by the MEDLINE indexer.

In the full-text database, we focused upon those fields unique to the full-text: the text (TX) and cited reference (CR) fields. Through a succession of Boolean operations we gathered answers to the following questions: 1) In how many of the items did the terms appear in the TX field? 2) Of those, how many could not have been retrieved from a search of the title and abstract? 3) How many of the records were retrieved only because of the presence of terms in the cited reference portion of the full-text record?

RESULTS

Table 1 shows the total items retrieved by the searches for each filter in each database. In all cases, the full-text search retrieved many more items than the MEDLINE bibliographic search.

Table 1
Items Retrieved by Bibliographic and Full-Text Versions of NEJM

	NEJM MEDLINE	NEJM Full-text
Randomized controlled trial	442	1377
Other trials	809	2268
Cohort studies	91	529
Followup studies	260	617
Multicenter studies	186	532

Results for the MEDLINE searches appear in Table 2. Total MEDLINE results for the various filters ranged from a low of 91 retrieved in the cohort filter to a high of 809 retrieved by the filter for prospective, controlled, and other clinical trials not

explicitly designated as randomized (labeled "other trials"). The middle column shows how many of the total items retrieved contained MeSH terms and/or publication types corresponding to the filter. The last column shows how many of those records did not also contain textwords describing the filter in the title or abstract. This number represents those MEDLINE records which were retrieved only because the indexer had assigned a MeSH term or publication type. In other words, if MEDLINE searchers had to rely only on the title and abstract for retrieval, all of these would have been missed.

Table 2
Records Retrieved by Value-Added Fields in Bibliographic Database

	Total MED-LINE	Value-added fields: DE,PT	Retrieved ONLY by Value-Added Fields
Randomized	442	418	80 (18%)
Other trials	809	775	470 (58%)
Cohort studies	91	47	10 (11%)
Followup studies	260	252	231 (89%)
Multicenter studies	186	182	118 (63%)

The first filter we tested was an attempt to identify those items which indicated the greatest methodological rigor, the randomized controlled trial. Searching all available fields in MEDLINE (TI,AB,DE,PT) retrieved 442 items. 418 of those included either the publication type RANDOMIZED CONTROLLED TRIAL, or the MeSH terms RANDOMIZED CONTROLLED TRIALS or RANDOM ALLOCATION. In 80 (18%) of those, neither the title nor the abstract mentioned the randomization. It was only because of the inclusion of the value-added, indexer-assigned fields that these 80 were retrievable at all.

The same protocol was followed for the other four filters. The value-added MEDLINE fields made a particularly large contribution to the searches for other trials, followup/longitudinal studies, and multicenter studies, where in each case over 60% of the studies could not have been recovered without the presence of the MeSH headings and publication types.

Table 3 examines the impact of unique fields only searchable in the full-text (text and cited reference) fields on retrieval. The middle column represents all items retrieved by searching the text field which did not also contain the search terms in the title or the abstract. For example, COHORT appeared in 529 full-text records; 372 (70%) of those were retrieved only because the TX field was searchable. In those 372, COHORT did not appear in the titles or abstracts. The right column lists items retrieved only because COHORT appeared in one or more cited references. That is, of the 529 records retreived by the term COHORT, 141 or 27% of those were retrieved solely because of the occurrence(s) in the cited references. While records retrieved by the cited references are less likely than those retrieved by the text to be valid instances of the study design, it is interest to note the number and percentage retrieved by this field.

Table 3
Records Retrieved by Value-Added Fields in Full-Text Database

	Total Full-text	TX not TI,AB	CR not TI,AB, TX
Randomized	1377	1079 (78%)	734 (53%)
Other trials	1786	1405 (79%)	460 (26%)
Cohort studies	529	372 (70%)	141 (27%)
Followup studies	617	408 (66%)	185 (30%)
Multicenter studies	532	247 (46%)	222 (42%)

DISCUSSION

With every filter tested the full-text database search yielded considerably more records than MEDLINE. This is to be expected, since the full-text provides so many more access points.

The high proportion of citations retrieved due to the availability of MeSH headings and publication types related to study design in MEDLINE suggests that the intellectual effort involved in indexing does indeed add value to the MEDLINE records, providing access points to research methodologies that would not otherwise be available. The data on followup studies and multicenter studies both suggest instances in which the indexer judged from the article that the study was one of these types. We know that authors do not always state that they are presenting a follow-up or a multicenter study. Instead they often list the number of years covered by the study of the names of the institutions where the study was conducted.

The text field accounted for the greatest number of unique items with each strategy tested in the full-text database. With the exception of MULTICENTER STUDIES, over 66% of the items retrieved from each of these searches were retrieved only because of the TX field. The terms did not appear in the titles or abstracts.

While the cited reference field yielded items not retrieved by the bibliographic record, it should be noted that those retrieved only from the cited reference field are least likely to be true examples of the study type, although that would have to be confirmed by further testing.

Although there were no relevance judgments available for these data, the sheer volume of additional records in the full-text database suggests that at least some of the additional records would indeed be examples of the study designs. In the first phase of the MEDLINE/Full-Text Research Project [9] we found that, on average, full-text searches resulted in .34 precision. Even if only one-third of the records retrieved in the full-text database are relevant, we can postulate that a sizeable number of additional relevant records would be retrieved for each study design.

ACKNOWLEDGMENT

This study was supported in part by NLM grant LM04605 and the searching was supported by Knight-Ridder Information, Inc. The opinions expressed are solely those of the authors.

REFERENCES

1. Kirplani H, Schmidt B, McKibbon KA et al: Searching MEDLINE for randomized clinical trialsinvolving care of the newborn. Pediatrics. 1989;83:543-6.

2. Dickersin K, Hewitt P, Mutch L, et al. Perusing the literature: comparion of MEDLINE searching with a perinatal trials database. Controlled Clinical Trials. 1985; 6:307-17.

3. Hewitt P, Dickersin K, Chalmers TC. More on MEDLINE searches. Controlled Clinical Trials. 1988; 9:85-90.

4. Poynard T, Conn HO. The retrieval of randomized clinical trials in liver diseases from the medical literature: a comparison of MEDLARS and manual methods. Controlled Clinical Trials. 1985; 6:271-9.

5. Herxheimer A. Randomised controlled trials: the Cochrane Collaboration. Journal of the Royal College of Physicians of London. 1993; 27:180.

6. Jadad AR, McQuay HJ. A high-yield strategy to identify randomized controlled trials for systematic reviews. Online Journal of Current Clinical Trials. 1993; 33.

7. Wilczynski NL, Walker CJ, McKibbon KA. Assessment of methodologic search filters in MEDLINE. Proceedings of the Annual Symposium for Computer Applications in Medicial Care. 1994; 17:601-5.

8. Miller C. Detecting duplicates: a searcher's dream come true. Online. 1990; 14: 27-34.

9. McKinin EJ, Sievert MC, Johnson ED, Mitchell JA.: The MEDLINE/Full-Text Research Project. Journal of the American Society for Information Science. 1991;42: 297-307.

APPENDIX

In the search strategies below the question mark indicates truncation; the (#n) operator means that the

terms must appear within # words of each other but may appear in either order; the (w) operator means that the words must appear next to each other in that order.

Randomized controlled trial;

randomi? or (random?? (5n) (allocat? or assign? or sampl???)) or dt=randomized controlled trial

N.b. incorporates the MeSH terms "random allocation" and "randomized controlled trials".

Other controlled studies:

placebo? or crossover or cross (w) over or prospective? or ((controlled or clinical) (w) trial? ?) or ((double or single or triple) (5n) (mask? or blind???)) or dt=clinical trial

N.b. incorporates the following MeSH terms: "placebo," "cross-over studies," "prospective studies," "clinical trials," "single-blind method," "double-blind method."

Cohort studies:

cohort

N.b. retrieves the MeSH term "cohort studies."

Multicenter studies:

(multicent? or multi()cent??? or multi()institution?) (3n) (stud??? or trial? ?) or dt=multicenter study

N.b. incorporates the MeSH term "multicenter studies."

Longitudinal studies:

(longitudinal or followup or follow) (5n) (stud??? or design)

N.b. incorporates the MeSH terms, "longitudinal studies" and "follow-up studies."